

SF1901: SANNOLIKHETSTEORI OCH
STATISTIK
FÖRELÄSNING 6
MER ON VÄNTEVÄRDE OCH VARIANS.
KOVARIANS OCH KORRELATION.
STORA TALENS LAG.

Tatjana Pavlenko

14 september 2015



PLAN FÖR DAGENS FÖRELÄSNING

- ▶ Repetition av begrepp oberoende s.v. (Kap. 4.5)
- ▶ Repetition av begrepp väntevärde och varians för diskreta och kontinuerliga s. v. Flerdimensionella s.v. (Kap. 5.1-5.2)
- ▶ Väntevärde för en funktion av flerdimensionell s.v. (Kap. 5.3, (c))
- ▶ Beroendemått: Kovarians och korrelation (Kap. 5.4, (c))
- ▶ Summa och linjärkombination: väntevärde och varians. (Kap. 5.5)
- ▶ Stora talens lag (Kap. 5.6)



OBEROENDE S.V. (REP.)

- ▶ Betrakta en tvådimensionell s.v. (X, Y) Intuitivt: man kan anse att X och Y är oberoende om *händelserna* $\{X \in A\}$ och $\{Y \in B\}$ är oberoende för alla mängder A och B . Detta leder oss till följande
- ▶ **Def:** De s.v. X och Y kallas *oberoende* om

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

för alla mängder A och B .

- ▶ **Sats:** De s.v X och Y är **oberoende** om och endast om

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{för alla } x \text{ och } y,$$

eller

$$p_{X,Y}(j, k) = p_X(j)p_Y(k) \quad \text{för alla } j \text{ och } k, \text{ för diskreta s.v.}$$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{för alla } x \text{ och } y, \text{ för kontinuerliga s.v.}$$



KTH Network

FÖRDELNING FÖR MAX OCH MIN (REP.)

- ▶ **Sats:** Låt X_1 och X_2 vara oberoende s.v. med fördelningsfunktioner $F_{X_1}(x)$ respektive $F_{X_2}(x)$. Definiera $U = \min(X_1, X_2)$ och $V = \max(X_1, X_2)$. Då gäller att

$$F_U(u) = 1 - (1 - F_{X_1}(u))(1 - F_{X_2}(u)),$$

$$F_V(v) = F_{X_1}(v)F_{X_2}(v).$$

- ▶ Sats kan vidare utvidgas för fler än två s.v.: Om X_1, \dots, X_n är oberoende och *lika fördelade med fördelningsfunktion* $F_X(x)$ så har $Y = \min_{i=1, \dots, n}(X_1, \dots, X_n)$ och $Z = \max_{i=1, \dots, n}(X_1, \dots, X_n)$ fördelningsfunktionerna

$$F_Y(y) = 1 - (1 - F_X(y))^n \text{ respektive } F_Z(z) = (F_X(z))^n.$$

- ▶ Bevis och exempel på tavlan.



FÖRDELNING FÖR MAX OCH MIN (REP.)

- ▶ Exempel: *Motor*. En motor upphör helt att fungera när *samtliga* 4 cylindrar gått sönder. Antag att cylindrarnas livslängder X_i , $i = 1, \dots, 4$ (i år) är oberoende och likafördelade $Exp(\lambda)$ där $\lambda = 1/7$, dvs $E(X_i) = 7$. Låt en s.v. T vara tid tills motorn helt upphör att fungera, då är $T = \max(X_1, \dots, X_4)$. Fördelningsfunktion för de enskilda cylindrarna är $F_{X_i}(x) = 1 - e^{-x/7}$ (samma för alla i), vilket ger

$$F_T(t) = F_{X_i}^4(t) = \left(1 - e^{-t/7}\right)^4.$$

- ▶ Om vi i stället är intresserade av tiden S då motor funktion blir nedsatt pga *någon* cylinder inte fungerar så får vi $S = \min(X_1, \dots, X_4)$ och fördelningsfunktionen för S blir

$$F_S(s) = 1 - (1 - F_{X_i}(s))^4 = 1 - (e^{-s/7})^4 = 1 - e^{-4s/7}.$$

- ▶ Minsta av n st. oberoende lika förd. s.v. med $Exp(\lambda)$ också är Exp -fördelad men $Exp(n\lambda)$!



- ▶ Väntevärde och varians: räkneregler (repetition på tavlan).
- ▶ Sats 5.1 från kap. 5.2 om väntevärde för en funktion av en s.v. kan utvidgas till en funktion av flera s.v.
- ▶ **Sats:** Låt (X, Y) vara en tvådimensionell s.v. och $g(X, Y)$ en reel funktion. Då gäller att

$$E(g(X, Y)) = \begin{cases} \sum_j \sum_k g(j, k) p_{X, Y}(j, k) & \text{eller} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X, Y}(x, y) \end{cases}$$

om (X, Y) är diskret respektive kontinuerlig.

- ▶ Intressanta fall är t ex

$$g(X, Y) = X + Y, \quad g(X, Y) = X \quad \text{sam} \quad t \quad g(X, Y) = XY.$$

- ▶ **Sats (väntevärde av en summa):** För godtyckliga s.v X och Y gäller att

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{X,Y}(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y) dx dy = \\ &= E(X) + E(Y). \end{aligned}$$

- ▶ Det diskreta fallet är helt analogt.
- ▶ $g(X, Y) = x$ (respektive $g(X, Y) = y$). Det gäller att

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) dx dy,$$

dvs man kan beräkna $E(X)$ och $E(Y)$ utan att först härleda de marginella fördelningarna!

- ▶ **Sats (5.3) (Linjäritet av väntevärde):** Om X och Y är s.v. med väntevärde $E(X)$ och $E(Y)$ och a, b, c är konstanter så gäller att

$$E(aX + bY + c) = aE(X) + bE(Y) + c.$$

- ▶ Obs! Satsen är helt generellt dvs gäller vid *godtyckliga beroende mellan X och Y .*
- ▶ **Sats (5.4):** Om X och Y är oberoende s.v. så gäller att

$$E(XY) = E(X)E(Y).$$

Detta utvidgas till fler än två oberoende s.v. X_1, \dots, X_n , se Sats 5.5

$$E(X_1 X_2 \dots X_n) = E(X_1)E(X_2) \dots E(X_n).$$



BEROENDEMÅTT

- ▶ När man behandlar tvådimensionella s.v är det viktigt att veta hur de två variablerna påverkar varandra, man brukar tala om *beroenden* mellan variabler. Hur beroende två s.v. X och Y är kan kvantifieras på olika sätt.
- ▶ Vi inför två närbesläktade beroendemått som beskriver graden av *linjärt* beroende/samband.
- ▶ Betrakta en tvådimensionell s.v. (X, Y) med ändliga väntevärden μ_X och μ_Y och stdavvikelser $D(X) = \sigma_X$ och $D(Y) = \sigma_Y$.



BEROENDEMÅTT: KOVARIANS OCH KORRELATION

- ▶ **Def:** Kovariansen mellan X och Y , betecknad med $C(X, Y)$, definieras som

$$C(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

- ▶ Tolkning av $C(X, Y)$. Samband mellan X och Y 's variation. Exempel på tavlan.
- ▶ Räkningregler för kovarianser: X, Y, Z är s.v., a, b, c, d konstanter.

$$C(X, X) = V(X), \quad C(X, Y) = C(Y, X),$$

$$C(aX + b, cY + d) = acC(X, Y),$$

$$C(X + Y, Z) = C(X, Z) + C(Y, Z).$$

- ▶ Allmänt: Kovariansen är bilinjär (dvs linjär i både argument).

$$C\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_i \sum_j a_i b_j C(X_i, Y_j).$$



- ▶ **Def:** Korrelationskoefficienten $\rho(X, Y)$ definieras som

$$\rho(X, Y) = \frac{C(X, Y)}{D(X)D(Y)} = \frac{C(X, Y)}{\sigma_X \sigma_Y}.$$

- ▶ Korrelationskoefficienten $\rho(X, Y)$ för en tvådimensionell s.v. är *dimensionslös* och satisfierar

$$-1 \leq \rho(X, Y) \leq 1.$$



- ▶ **Stas (5.8), beräkning a kovarians:** Följande samband gäller

$$C(X, Y) = E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y.$$

- ▶ **Sats:** Om X och Y är oberoende så är de också okorrelerade.
Bevis:

$$E(XY) = E(X)E(Y) \text{ för ober. } X, Y \Rightarrow C(X, Y) = 0.$$

- ▶ **Omvändningen till satsen gäller inte!** Okorrelerade s.v är inte nödvändigtvis oberoende. Exempel på tavlan.
- ▶ Kovariansen är viktig för att förstå variation i *summor* av s.v.

LINJÄRKOMBINATION

- ▶ Kovariansen är viktig för att förstå *variation* i summor av s.v.
- ▶ Sats 5.3 kan utvidgas till fler än två s.v. Låt oss se på linjärkombination

$$a_1X_1 + a_2X_2 + \dots + a_nX_n + b$$

av n s.v. X_1, \dots, X_n , a_1, a_2, \dots, a_n , b är konstanterna.

- ▶ **Sats (5.11):** För *alla* X_1, \dots, X_n gäller

$$E\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i E(X_i) + b,$$

$$V\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{1 \leq j < k \leq n} a_j a_k C(X_j, X_k).$$

- ▶ Ex på tavla!



- ▶ **Följdsats (5.11.1):** För *oberoende* X_1, \dots, X_n gäller

$$E\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i E(X_i) + b,$$

$$V\left(\sum_{i=1}^n a_i X_i + b\right) = \sum_{i=1}^n a_i^2 V(X_i).$$

- ▶ **Följdsats (5.11.3):** Om X_1, \dots, X_n är oberoende s.v, var och en med väntevärdet μ och std σ , och

$$\bar{X} = \sum_{i=1}^n (X_i + \dots X_n) / n$$

är deras aritmetiska medelvärde, så gäller att

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \sigma^2 / n, \quad D(\bar{X}) = \sigma / n.$$

STORA TALENS LAG.

- ▶ Stora talens lag är ett av de viktigaste resultaten inom sannolikhets teorin.
- ▶ Denna lag, först formulerad av den schweiziske matematikern Jacob Bernoulli (1654–1705), utsäger att *aritmetiska medelvärdet av flera oberoende s.v. med samma väntevärde μ ligger nära μ* , bara antalet är tillräckligt stort.





FIGUR : Jacob Bernoulli grundlade sannolikhetsläran med den postumt utgivna *Ars Conjectandi* (*Konsten att gissa* (1713)) där de stora talens lag presenteras för första gången.

- ▶ **Sats:** Låt X_1, X_2, \dots vara en följd av oberoende (likafördelade) s.v., alla med samma väntevärde $E(X_i) = \mu$ och std $D(X_i) = \sigma < \infty$.
Låt

$$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$$

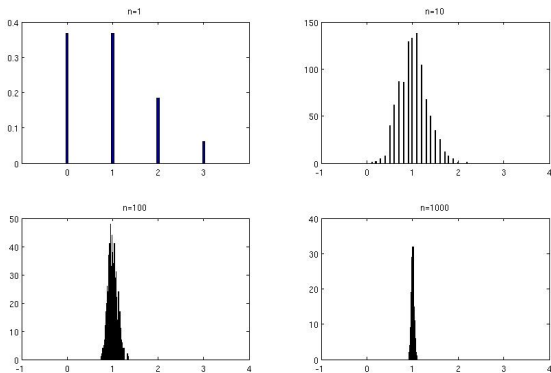
vara medelvärdet av de n första variablerna. Då gäller, för godtyckligt $\varepsilon > 0$, att

$$P(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0 \text{ då } n \rightarrow \infty.$$

- ▶ Detta säger att s.v \bar{X}_n , bestående av medelvärdet av de n s.v. X_1, \dots, X_n har en fördelning som koncentrerar sig runt $\mu = E(X_i)$.



STORA TALENS LAG (FORTS.)



FIGUR : Sannolikhetsfunktionen för medelvärdet $\bar{X}_n = (X_1 + \dots + X_n)/n$, där de enskilda observationerna är $Po(1)$, för $n = 1, 10, 100$ och 1000 . Den diskreta fördelningen koncentreras alltmer runt värdet 1, dvs väntevärdet.

STORA TALENS LAG (FORTS.)

- ▶ *Tolkning:* medelvärde är en bra uppskattning av väntevärde!
- ▶ *Stora talens lag är en av grundstenarna inom empirisk vetenskap.* Om man gör många oberoende observationer av någon s.v., t ex hållfastheten hos en metall, blodtrycket hos patienter behandlade med en ny medicin, eller livslängden hos personer i ett försäkringskollektiv. Då kommer medelvärdena av dessa observationer att ligga nära det *sanna* väntevärdet.
- ▶ *Om vi inte vet det sanna väntevärdet kan vi gissa, eller skatta väntevärdet med medelvärdet.* Detta förhållande är en viktig ingrediens i statistik delen av kursen.

